CLAIMS

1. A method for deriving server resource utilization estimates for a server cluster, the method comprising:

recording server cluster data during operation of the server cluster, at least some of the server cluster data indicating server resource parameter values;

using a load simulation tool that, using the recorded data, determines a maximum load that can be handled by the server cluster;

specifying a load to be handled by the server cluster; and

deriving server resource utilization estimates corresponding to the specified load.

2. The method as recited in claim 1, further comprising:

displaying the server resource utilization estimates; and

recommending a plan to optimize processing of the specified load.

3. The method as recited in claim 2, wherein the plan recommends a change in the hardware configuration of the server cluster.

4. The method as recited in claim 1, wherein the maximum load, the recorded values, the specified load, and the server resource utilization estimates are stored in non-volatile memory.

5. The method as recited in claim 1, wherein the using a load simulation tool comprises:

creating a test script from the recorded values;

running the test script on a master client to simulate load and server resource utilization conditions that existed on a server when the recorded values were recorded; and

increasing the load on the server, when the test script is running, until a maximum load that can be handled by the server is obtained.

6. The method as recited in claim 5, wherein the increasing the load on the server further comprises multiplying the number of users utilizing the server cluster when the recorded values were recorded, thereby multiplying the resources utilized by the users.

7. The method as recited in claim 5, wherein the increasing the load on the server further comprises decreasing the amount of time between user requests, thereby increasing the resources utilized by the users.

8. The method as recited in claim 5, wherein the increasing the load on the server until a maximum load that can be handled by the server is obtained, further comprises:

observing a service rate exhibited by the server on which the simulation is being performed; and

recognizing that the maximum load has been obtained when an increase in the load does not increase the service rate.

9. The method as recited in claim 5, wherein:

the server cluster contains a set of identical servers;

running the test script run on the master client simulates server cluster operation on only one of the servers of the server cluster; and

the method further comprises extrapolating the results obtained on the one server using the number of servers in the set of identical servers to obtain the maximum load that can be handled by the server cluster.

10. The method as recited in claim 5, wherein:

the server cluster contains a set of non-identical servers;

running the test script run on the master client further comprises running the test script on each of the non-identical servers in the server cluster; and

the method further comprises summing the results obtained from each non-identical server in the cluster to obtain the maximum load that can be handled by the server cluster.

11. The method as recited in claim 1, wherein the recording server cluster data during operation of the server cluster comprises recording data directly from the server cluster and recording data that is input by a server cluster user.

12. The method as recited in claim 1, wherein the server resource utilization estimates comprise estimates for one or more of the following: processor utilization, memory utilization, communication bandwidth utilization, and general server utilization.

13. The method as recited in claim 1, wherein the server resource utilization comprises general server utilization, the method further comprising:

Deriving general server utilization by solving:

$$U = \frac{L}{X}$$

Wherein U is the general server utilization; X is the maximum load that can be handled by the server cluster; and L is the specified load.

14. The method as recited in claim 1, wherein the server resource utilization comprises processor utilization, the method further comprising:

deriving processor utilization by solving:

$$U_{CPU}=\frac{a}{e^{b\cdot L}}$$

wherein U_{CPU} is processor utilization; L is the specified load; a is processor regression constant a; and b is processor regression constant b.

15. The method as recited in claim 1, wherein the server resource utilization comprises communication bandwidth utilization, the method further comprising:

deriving communication bandwidth utilization by solving:

$$U_B = \frac{F_{TCP}}{B}\cdot(c + d\cdot L)$$

wherein U_B is communication bandwidth utilization; L is the specified load; c is processor regression constant c; d is processor regression constant d; F_{TCP} is a transmission overhead factor; and B is the total communication bandwidth available.

16. The method as recited in claim 1, wherein the server resource utilization comprises memory utilization, the method further comprising:

deriving memory utilization by solving:

$$U_M = \frac{N \cdot (M_{TCP} + M_{IISStruct}) + M_{OS} + M_{IIS}}{M}$$

wherein N is a total number of concurrent connections derived by solving:

$$N = \frac{L}{(X - L)} + S1 \cdot L$$

wherein: U_M is memory utilization; M_{TCP} is a an amount of memory necessary to support the connections for communications; $M_{IISStruct}$ is the amount of memory necessary to support data structures associated with each connection; M_{OS} is the amount of memory required by a server operating system; M_{IIS} is the amount of memory required by a server communication program; M is the total amount of memory available; L is the specified load; X is the maximum load that can be handled by the server cluster; and $S1$ is a connection memory factor that is the adjusted average of the incoming connections at different speeds.

17. The method as recited in claim 1, wherein the server resource utilization comprises general server utilization, the method further comprising:

18. The method as recited in claim 1, wherein the server resource utilization comprises processor utilization, the method further comprising:

deriving processor utilization by solving:

$$U_{CPU}=\frac{a}{e^{b\cdot L}}$$

wherein U_{CPU} is processor utilization; L is the specified load; a is processor regression constant a; and b is processor regression constant b.

19. The method as recited in claim 1, wherein the server resource utilization comprises communication bandwidth utilization, the method further comprising:

deriving communication bandwidth utilization by solving:

$$U_B = \frac{F_{TCP}}{B} \cdot (c + d \cdot L)$$

wherein U_B is communication bandwidth utilization; L is the specified load; c is processor regression constant c; d is processor regression constant d; F_{TCP} is a transmission overhead factor; and B is the total communication bandwidth available.

20. The method as recited in claim 1, wherein the server resource utilization comprises memory utilization, the method further comprising:

deriving memory utilization by solving:

$$U_M = \frac{N \cdot (M_{TCP} + M_{IISStruct}) + M_{OS} + M_{IIS}}{M}$$

wherein N is a total number of concurrent connections derived by solving:

$$N = \frac{L}{(X - L)} + S1 \cdot L$$

wherein: U_M is memory utilization; M_{TCP} is a an amount of memory necessary to support the connections for communications; $M_{IISStruct}$ is the amount of memory necessary to support data structures associated with each connection; M_{OS} is the amount of memory required by a server operating system; M_{IIS} is the amount of memory required by a server communication program; M is the total amount of memory available; L is the specified load; X is the maximum load that can be handled by the server cluster; and $S1$ is a connection memory factor that is the adjusted average of the incoming connections at different speeds.

21. The method recited in claim 1, wherein the server resource utilization is processor utilization, the method further comprising:

finding a functional dependency approximation between processor utilization and load;

transforming functional dependency into linear form by using logarithmic transformation;

deriving first and second processor regression constants using linear regression methodology;

dividing the first processor regression constant by *e* to the power of the product of the second processor regression constant and the specified load to obtain the processor utilization estimate.

22. The method as recited in claim 1, wherein the server resource utilization is communication bandwidth utilization, the method further comprising:

finding a functional dependency approximation between communication bandwidth utilization;

transforming functional dependency into linear form by using logarithmic transformation;

deriving first and second bandwidth regression constants using linear regression methodology;

deriving a transmission overhead factor that, when applied to a certain size web page, results in the actual capacity necessary to transmit the web page;

deriving a weighted communication overhead factor by dividing the transmission overhead factor by the available communication bandwidth;

deriving an adjusted communication load by adding the first bandwidth regression constant to the product of the specified load and the second bandwidth regression constant; and

determining the communication bandwidth utilization estimate by multiplying the weighted communication overhead factor by the adjusted communication load.

23. The method as recited in claim 1, wherein the server resource utilization is memory utilization, the method further comprising:

deriving a connection memory factor that is the adjusted average of the incoming connections at different speeds;

deriving a weighted connection memory factor by multiplying the connection memory factor by the specified load;

deriving a page load ratio by dividing the specified load by the difference of the maximum load value and the specified load;

deriving a total number of concurrent connections by adding the weighted connection memory factor and the page load ratio; and

deriving a gross memory utilization by multiplying the total number of concurrent connections by the sum of the amount of memory necessary to support each connection for communications and the amount of memory necessary to support data structures associated with each connection, and adding the amount of memory required by a server operating system and the amount of memory required by the server communication program; and

deriving the memory utilization estimate by dividing the gross memory utilization by total memory available.

24. The method as recited in claim 1, wherein the server resource utilization is general server utilization, the method further comprising:

dividing the specified load by the maximum load to derive the general server utilization estimate.

25. One or more computer-readable media having computer-readable instructions thereon which, when executed by one or more computers, cause the computers to implement the method of claim 1.

26. A simulation tool for use in determining server resource utilization estimates in a server cluster having one or more servers, the load simulation tool comprising:

a user interface configured to receive data input from a user;

at least one filter or monitor configured to record operational data from one or more of the servers in the server cluster;

the simulation tool being configured to create a test script from the recorded data and the received data, and to run the test script from a master client connected to the server cluster to simulate load and other server conditions that existed when the operational data was recorded; and

the user interface being further configured to display utilization of server resources during the running of the test script.

27. The simulation tool as recited in claim 26, wherein the simulation tool being configured to create the test script is further configured to allow the user to create the test script to increase the load on the server on which the simulation is running and observe the effect of such an increase in load on the server resource utilization displays.

28. A system, comprising:

a server cluster having one or more servers, one of which is a primary server that controls the operation of the server cluster;

a cluster controller resident in memory on the primary server of the server cluster, the cluster controller controlling communications between the primary server and secondary servers, if any, and between clients and the server cluster;

an operating system resident in the memory of the primary server;

a communications program within the cluster controller to provide communications capability for the system;

a filter to collect server data indicating certain operating parameters for the server cluster;

a monitor on each server in the server cluster to collect server data indicating certain operating parameters for the server cluster;

a user interface to collect data input by a user;

a capacity planner within the cluster controller configured to utilize the collected data to derive one or more server resource utilization estimates for server resources to determine how handling a specified load will affect the utilization of the server resources, and to produce a plan recommending changes to be made to the server cluster to adequately accommodate the specified load; and

a load simulation tool configured to use the collected data to create a simulation script that, when run on a master client, simulates the operation of the server cluster system to allow the user to find the maximum load that the server cluster can handle; and

wherein the maximum load obtained through the use of the load simulation tool is utilized in the derivation of the one or more server resource utilization estimates.

29. The system as recited in claim 28, wherein the filter is an ISAPI filter.

30. The system as recited in claim 28, wherein the collected data and the plans are stored in the memory.

31. The system as recited in claim 28, wherein the simulation script is run from a master client connected to the server cluster, and wherein the simulation is performed on only one server of the server cluster.

32. The system as recited in claim 31, wherein the load simulation tool is further configured to extrapolate results from the simulation on one server in the server cluster to derive results for the total number of servers in the server cluster.

33. The system as recited in claim 28, wherein:

the load simulation tool is further configured to run a simulation script from a master client connected to the server cluster;

the simulation is performed on each server in the server cluster; and

results from each server are summed to derive results of the total number of servers in the server cluster.

34. The system as recited in claim 28, wherein the server resource utilization derived by the capacity planner comprises general server utilization, and the capacity planner is further configured to derive general server utilization by solving:

$$U = \frac{L}{X}$$

wherein U is the general server utilization; X is the maximum load that can be handled by the server cluster which is determined by the load simulation tool; and L is the specified load.

35. The system as recited in claim 28, wherein the server resource utilization derived by the capacity planner comprises general server utilization, and the capacity planner is further configured to derive general server utilization by solving:

$$U_{CPU}=\frac{a}{e^{b\cdot L}}$$

wherein U_{CPU} is processor utilization; L is the specified load; a is processor regression constant a; and b is processor regression constant b.

36. The system as recited in claim 28, wherein the server resource utilization derived by the capacity planner comprises communication bandwidth utilization, and the capacity planner is further configured to derive communication bandwidth utilization by solving:

$$U_B = \frac{F_{TCP}}{B}\cdot(c + d\cdot L)$$

wherein U_B is communication bandwidth utilization; L is the specified load; c is processor regression constant c; d is processor regression constant d; F_{TCP} is a transmission overhead factor; and B is the total communication bandwidth available.

37. The system as recited in claim 28, wherein the server resource utilization derived by the capacity planner comprises communication bandwidth utilization, and the capacity planner is further configured to derive communication bandwidth utilization by solving:

$$U_M = \frac{N \cdot (M_{TCP} + M_{IISStruct}) + M_{OS} + M_{IIS}}{M}$$

wherein N is a total number of concurrent connections derived by solving:

$$N = \frac{L}{(X - L)} + S1 \cdot L$$

wherein: U_M is memory utilization; M_{TCP} is a an amount of memory necessary to support the connections for communications; $M_{IISStruct}$ is the amount of memory necessary to support data structures associated with each connection; M_{OS} is the amount of memory required by a server operating system; M_{IIS} is the amount of memory required by a server communication program; M is the total amount of memory available; L is the specified load; X is the maximum load that can be handled by the server cluster; and $S1$ is a connection memory factor that is the adjusted average of the incoming connections at different speeds.